Stansbury & Janik (2019) *Current Biology* 29: xxx-xxx, DOI 10.1016/j.cub.2019.05.071

Formant Modification through Vocal Production Learning in Grey Seals

Amanda L. Stansbury[1,†] and Vincent M. Janik[1,2,*]
[1] Scottish Oceans Institute, School of Biology, University of St Andrews, Fife KY16 8LB, UK
[2] Lead contact
* Correspondence: vj@st-andrews.ac.uk
† Current address: El Paso Zoo, El Paso, Texas, USA

SUMMARY

Vocal production learning is a rare communication skill and has only been found in selected avian and mammalian species [1-4]. While humans use learned formants and voiceless sounds to encode most lexical information [5], evidence for vocal learning in other animals tends to focus on the modulation pattern of the fundamental frequency [3, 4]. Attempts to teach mammals to produce human speech sounds have largely been unsuccessful, most notably in extensive studies on great apes [5]. The limited evidence for formant copying in mammals raises the question whether advanced learned control over formant production is uniquely human. We show that grey seals (*Halichoerus grypus*) have the ability to match modulations in peak frequency patterns of call sequences or melodies by modifying the formants in their own calls, moving outside of their normal repertoire's distribution of frequencies and even copying human vowel sounds. Seals also demonstrated enhanced auditory memory for call sequences by accurately copying sequential changes in peak frequency and the number of calls played to them. Our results demonstrate that formants can be influenced by vocal production learning in non-human vocal learners, providing a mammalian substrate for the evolution of flexible information coding in formants as found in human language.

RESULTS AND DISCUSSION

Most animals are born with a relatively fixed set of vocalizations to use in their acoustic communication. Some species can enlarge their repertoires over time by either combining signal elements into new sequences (usage learning) or through vocal production learning, the skill to use previous experience with model sounds to either copy them or to produce novel sounds [1]. Humans used these skills to evolve a highly flexible and versatile acoustic communication system that enables us to speak about novel objects and contexts by expanding our repertoire. While vocal learning does not necessarily lead to language evolution, it is a key pre-requisite for spoken language. Therefore, scientists have studied primate vocal skills in great detail searching for vocal learning as one of the precursors to language evolution. Interestingly, the only evidence for vocal production learning in nonhuman primates concerns either non-vocal sound production, like in lip smacking [6] and whistling [7], or subtle changes in parameters [8, 9] that are far from the frequency modulations that humans can achieve. Recently, accelerated vocal development in primates has also been described as a form of vocal learning [10]. All of the call modifications found in primates, however, are changes to existing call categories rather than resulting in novel signals, and it is often questionable whether the animals copied a novel sound or just used a version of a call that was already in their repertoire [11]. In contrast, other taxa can use learning to modify their signals much more substantially (e.g. songbirds [3], parrots [12, 13], bats [14, 15], cetaceans [16-18], elephants [19]). Some song birds even copy alarm calls and song elements of other species [20]. Similarly, intelligible copies of human speech sounds have been reported for parrots [21], song birds [22, 23], elephants [24] and seals [25].

Most evidence for vocal learning shows flexibility in modifying the fundamental frequency of sounds [3, 4]. This can be achieved by adjusting the tension of vocal cords and subglottal pressure. However, the main differences in voiced human speech sounds are caused by resonances in the supra-laryngeal tract that determine the energy content of harmonics in a sound. The resulting emphasized frequency bands are referred to as formants, which in humans differ between speech sounds, such as vowels. Formants have been reported in species-specific vocalizations of a variety of animals including birds and mammals [26], and most recently in alligators [27]. In most animals, formant spacing directly relates to body size and is used by conspecifics as an indicator of the latter [28]. Yet, in some cases formants change with behavioral context and can encode additional information. Diana monkeys [29] and meerkats [30] use formants in this way in their alarm calls and dogs shift formants down when defending food [31]. Rhesus macaques respond spontaneously to changes in formants [32] and also use formants to detect indexical cues [33] suggesting that they carry important information in primate communication in general.

How animals learn to modify formants has been studied in birds [22] but there are no such studies on mammals. Pinnipeds have recently been championed as a key study group in the field of human speech evolution because their skills may create a bridge between data on song birds and primates by showing vocal flexibility while using a production apparatus that is homologous to that of humans [34]. To investigate how flexible seals are in their use of formants, we studied how pinnipeds

copy sounds by training three juvenile grey seals (*Halichoerus grypus*) to imitate a variety of sound stimuli to systematically assess their vocal learning abilities. Grey seals produce moan calls in mother-pup interactions and in howling choruses on haul-out sites [35], contexts in which vocal learning can be advantageous [4].

In the first experiment, we transposed the frequency spectrum of seal moan calls (Figure S2) to create a sequence of moan calls with peak frequencies that were one to two standard deviations above or below the mean peak frequency of natural calls (1015 Hz ± 89 Hz SD, n=100) aiming to produce intervals based on musical scales. The fundamental frequency and other harmonics were changed accordingly to maintain the harmonic structure of the call. We then trained seal A to copy the call sequence, rewarding the animal not for exact matching but for cases in which the direction of peak frequency shifts between successive elements of the sequence matched the model. When the animal reached a criterion of 80% correct responses across 7 consecutive sessions in this training (which occurred after 1576 trails spread across 37 sessions), we started tests. While during training animal A was only exposed to up to three frequencies (880, 1046, and 1174 Hz, i.e. musical notes A5, C6 and D6), test stimuli extended over an octave (ranging from 698 to 2093 Hz, corresponding to musical notes F5 to C7). Throughout testing 15 sequences consisting of different combinations of up to ten of these notes were used. We used truncated as well as full 10 note sequences, resulting in a total of 92 test sequences. Rewards were again based on relative frequency shifts between successive elements of the sequence, not on absolute frequency matches. The seal matched the number of elements in a sequence played and absolute signal parameters

accurately (Figure 1A) even though only relative frequency changes in the right direction were required for a reward (see Videos S1 & S2 for examples, Table S1 for parameter descriptions). However, some parameters were matched better than others (Table 1). As was to be expected because of how we rewarded the animal, animal A matched the peak frequency and change in peak frequency of the signals (Figure 2A) but did not match the fundamental or change in fundamental frequency. Animal A also matched relative changes in frequency reliably but did not match absolute frequency well. Seals may be able to produce more accurate copies if reinforcement included absolute frequency matching as a criterion.

Our animal was also able to match combinations of notes up to the maximum tested, which was ten calls. The working memory for unrelated items in human and non-human primates is limited to a small number of 4-7 [36]. However, this number increases to around 10 in human memory for melodies [37]. Our results could indicate that seals either have improved working memory or that they perceived our modified call sequences as melodies.

Having demonstrated that a seal can learn to shift the peak frequency in its calls, we proceeded to test responses to vowels. We trained two new animals (animals B and C) in the same way as animal A using the same frequency-shifted moan calls. When they achieved our criterion of 80% correct responses across 7 consecutive sessions in matching sequences with three different notes (which occurred after 619 trails spread across 25 sessions for animal B and in 726 trails over 24 sessions for animal C), new training stimuli were brought into these sequences. These were five cardinal

vowels ([a], [e], [i], [ɔ], and [u]) as spoken by a human native English speaker with their frequency spectrum transposed so that they had a mean fundamental frequency similar to each seal's natural moan calls (Figure S1A). Seals were trained to copy a fixed sequence of these 5 vowels until they reached 80% correct responses across 7 consecutive sessions (which occurred after 623 trails spread across 28 sessions for animal B and in 569 trails over 27 sessions for animal C). For conducting formant measurements, we tested our seals with sequences of up to three of these vowels in random order. We used a total of 155 different sequences in these tests and tested each of these three times with each animal.

Both animals produced vowel sounds that could easily be identified by human listeners at the end of training (Figure 1B, Videos S3 and S4 for examples). During testing, we found that animals copied different aspects of the frequency spectrum to achieve this (Figure S1). Animal B changed the second and third formant to approximate the model, as well as the difference between formants (Figure 2B), but did not match the first formant. Animal C had a relatively constant first formant that did not change much but the seal changed the second formant to approximate the model and the difference between formants in the model sound (Figure 2C). While the seals were played each test signal multiple times, they performed comparably well in response to the first presentation of each stimulus since they had already been trained to produce these vowels in the training sessions (Table 1).

To demonstrate vocal learning in production rather than usage, it is imperative to demonstrate that learned sounds were not in the animal's repertoire before training

and also that the novel sounds are not normal part of the species repertoire. We use peak frequency here to demonstrate that the animals indeed learned to produce novel sounds. Using Levene's test, the peak and fundamental frequencies from the first 250 calls recorded at the start of training were compared to the last 250 calls recorded from each seal (Figures 3B, 3C and 3D). For all three seals, peak frequency variance significantly increased after training and included numerous cases of peak frequencies that could not be found in the pre-training repertoire. Increased peak frequencies were also higher than those found in naïve wild seals (Figure 3A). The variability of the fundamental frequency did not change with training for Animal A or B, but variability significantly decreased for animal C, and the trained animals showed less variation in the fundamental frequency than observed in wild grey seals.

Our grey seals demonstrated vocal production learning by producing calls that were novel to them and were not part of the normal grey seal vocal repertoire. This was suggested by an anecdotal report of a harbor seal called Hoover that produced human vowel sounds [25]. The report did not provide information on Hoover's pre-exposure repertoire or how his vowel sounds compared to the repertoire of naive harbor seals. While Hoover clearly mimicked human speech sounds, it was unclear whether this was achieved by stringing sounds from his pre-existing repertoire together [38], which would be a case of usage learning [1], or by creating novel sounds through production learning. Our data show that phocid seals can use vocal production learning to achieve this kind of formant copying. Hoover and our three grey seals were all juveniles when they learned their calls. This could be important

given that sensitive learning periods are not uncommon in vocal learners. Future studies need to address the potential for vocal learning in adulthood.

One other case of formant learning has been reported in an Asian elephant, inventively using its trunk to modify its mouth cavity to produce human vowels [24]. Wild elephants do not seem to use their trunk for call modification in this way [24]. This case is an impressive demonstration of trunk motor learning abilities but does not relate to learning abilities of the vocal tract itself. Other reports of speech-producing mammals only showed matches in duration, rhythm and frequency band used and mostly did not provide an adequate comparison with the species' natural repertoire [39-42]. These other examples come from toothed whales which are apt vocal learners [43] but whose vocal apparatus is so different from that of other mammals [44] that formant copying is challenging. All three of our seals also successfully matched varying sequences of sounds, confirming previous studies that demonstrated advanced usage learning [45, 46], sometimes called program-level imitation [47]. A recent study on zebra finches has highlighted how such usage learning complements production learning in bird song learning and human language acquisition [48]. Seals are an interesting model species for the comparison of vocal learning skills between humans and other mammals, since the anatomy and innervation of the vocal tract are highly similar [34]. Elephant trunk use, nasal sound production in cetaceans or the use of a syrinx and beak in birds involve muscles that are different to those used by humans for sound modifications in the larynx and supra-laryngeal structures. Future studies should focus on exploring the exact limits

of such copying skills in seals to see whether some modifications are easier to achieve than others.

Our sample size and experimental approach with artificially modified model sounds was suitable to explore pinniped vocal skills but does not inform us on how animals use these skills in the wild. One major context in which in-air vocalizations are used in wild pinnipeds is for mother-pup recognition during the nursing period. A wide variety of species have been found to use calls to facilitate reunions between mothers and their offspring [49]. In Northern fur seals, mothers remember the calls of their pups over periods of at least four years [50]. It is possible that learning allows animals to create novel calls by avoiding those they hear [1], and thereby being more recognizable in the presence of masking calls from other pups. It appears that bottlenose dolphins deploy their vocal learning skills in this way [16]. Another context in which vocal learning could be useful is in the production of vocal displays to attract mating partners or deter competitors. Several seal species (e.g. harbor seals [51], Weddell seals [52], bearded seals [53], walrus [54]) produce such displays. If vocal learning influences call diversity, this could be used as a measure of fitness as has been found in several song birds [3].

Alternatively, the learned control over muscles in the mouth and other parts of the vocal apparatus might be a by-product of adaptations of the respiratory tract for diving and suction feeding [55]. In that case, we would not expect an adaptive value of vocal learning per se and would expect learning to play no major role in pinniped communication. However, first evidence from male Southern elephant seals

suggests that they learn temporal features of dominance calls from the most successful males [56], suggesting that there is a benefit to vocal learning in at least this species.

Subtle changes of individual parameters of calls in mammals appear to be mediated by a mechanism different from the one that allows some species to produce completely novel calls [1, 2]. For non-human mammals, our results show that such an ability to produce novel signals does not only exist for modulation patterns of the fundamental frequency but also for the learning of formants, which are the main carriers of information in human language. Thus, the voluntary control of the mouth and laryngeal tract responsible for learned adjustments in spectral components such as formants is not unique to humans and may have a wider role in the flexibility of mammalian communication systems. While previous work on nonhuman primates has had little success in showing such vocal flexibility [2, 4], formant modulation is an area that deserves further investigation in a range of species to assess its role in the evolution of complex communication.

ACKNOWLEDGMENTS

We would like to thank the animal care staff of the Scottish Oceans Institute, especially Ryan Milne and Alicia Widmer. Kerri Rodriguez and Mafalda de Freitas helped with animal training. This research was conducted under Home Office license number 60/3303.

AUTHORS CONTRIBUTIONS

Both authors designed the study and wrote the manuscript. AS performed the experiments and analysis. VMJ secured research funding and supervised the research.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Stansbury & Janik (2019) *Current Biology* 29: xxx-xxx, DOI 10.1016/j.cub.2019.05.071

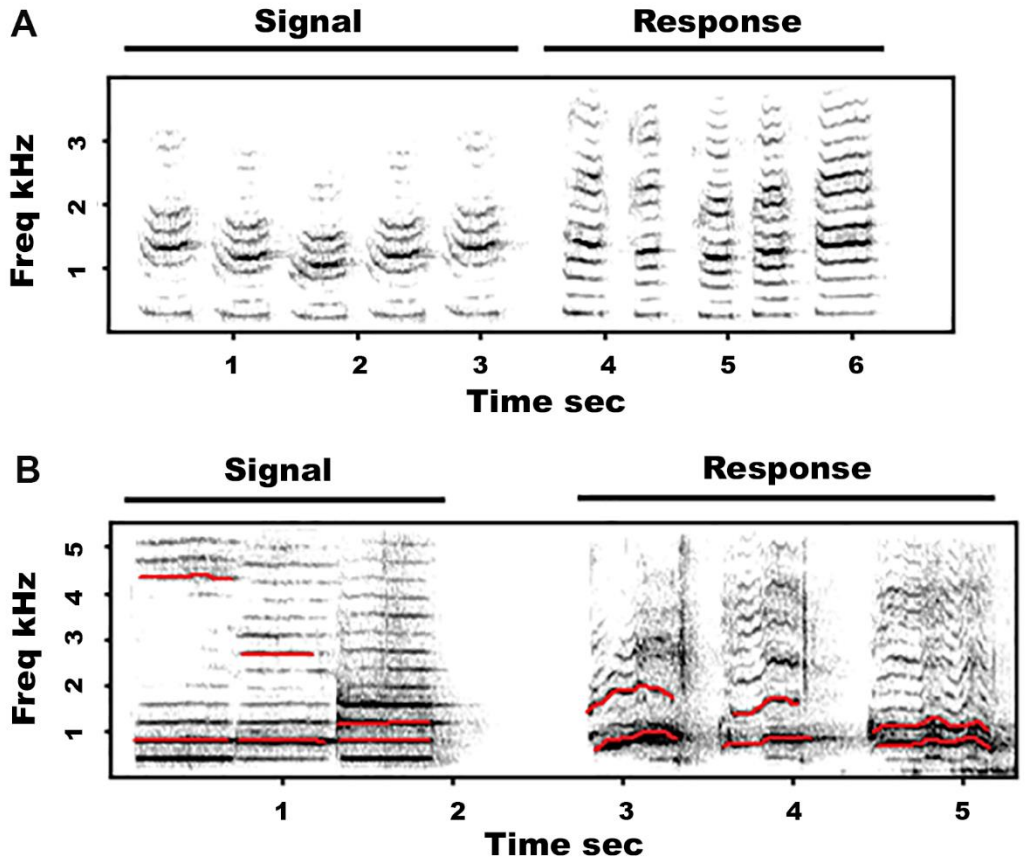

Figure 1. Spectrograms of two test trials showing the signal followed by the seal's response.

(A) A five note model signal and animal A's response. Audio included in Video S1.

(B) A three vowel model signal ([i], [e], and [ɑ]) and animal C's response. Formant 1 and 2 as determined by PRAAT are shown in red.

See also Figures S1 and S2 and Videos S3 and S4.

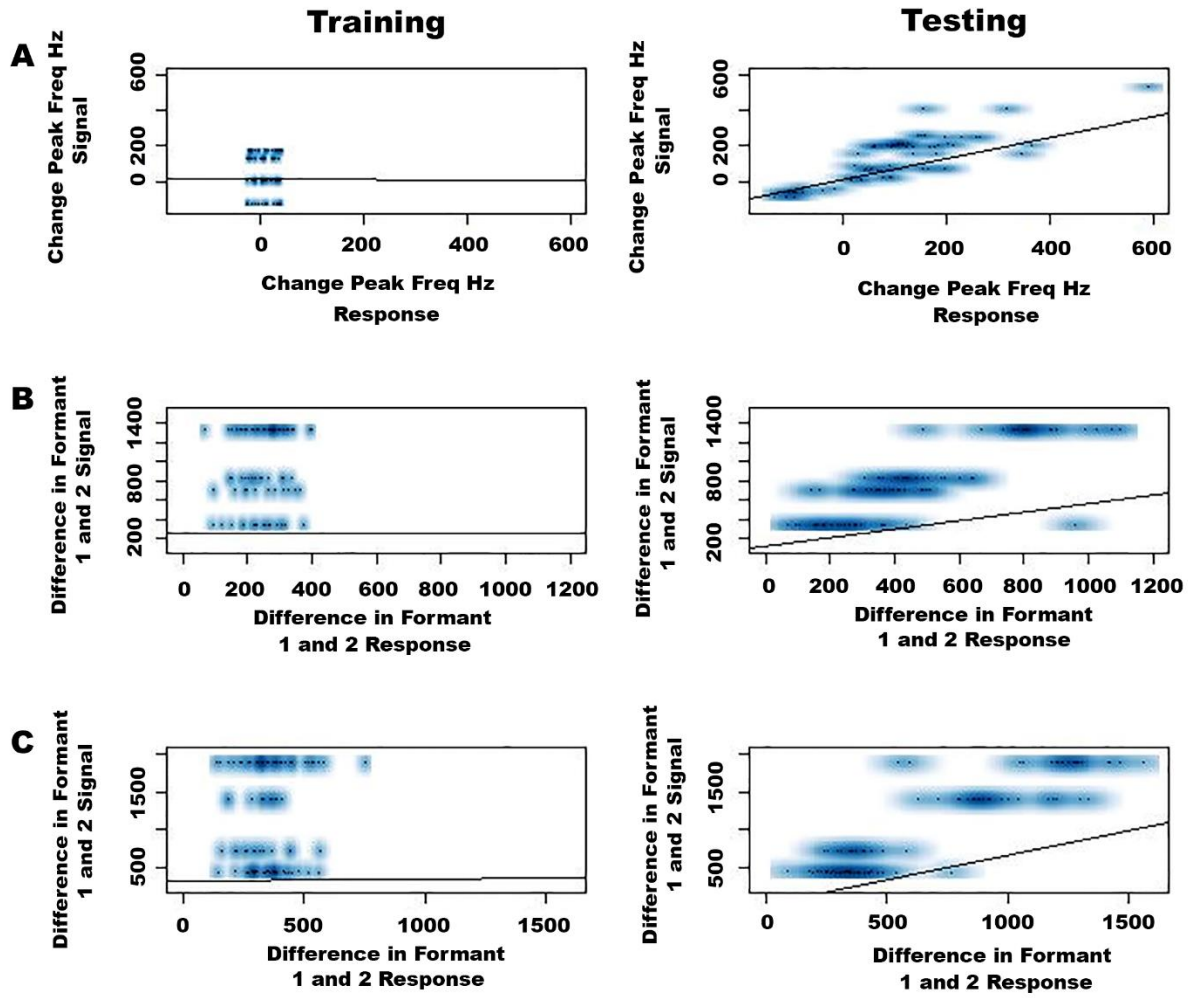

Figure 2. Scatter smooth plot comparing the signal to the seal's response at the start of training and at testing.

If there was no relationship between the signal and response (i.e. the animal did not match the signal) the regression line is flat with no slope. If there was a relationship between the signal and response (i.e. the animal did match the signal), the regression line has a positive slope (with a perfect match occurring with a 45 degree slope and equal axis values). See also Table S1 and Figure S2.

(A) Animal A's change in peak frequency

(B) Animal B difference between formant 1 and 2.

(C) Animal C difference between formant 1 and 2.

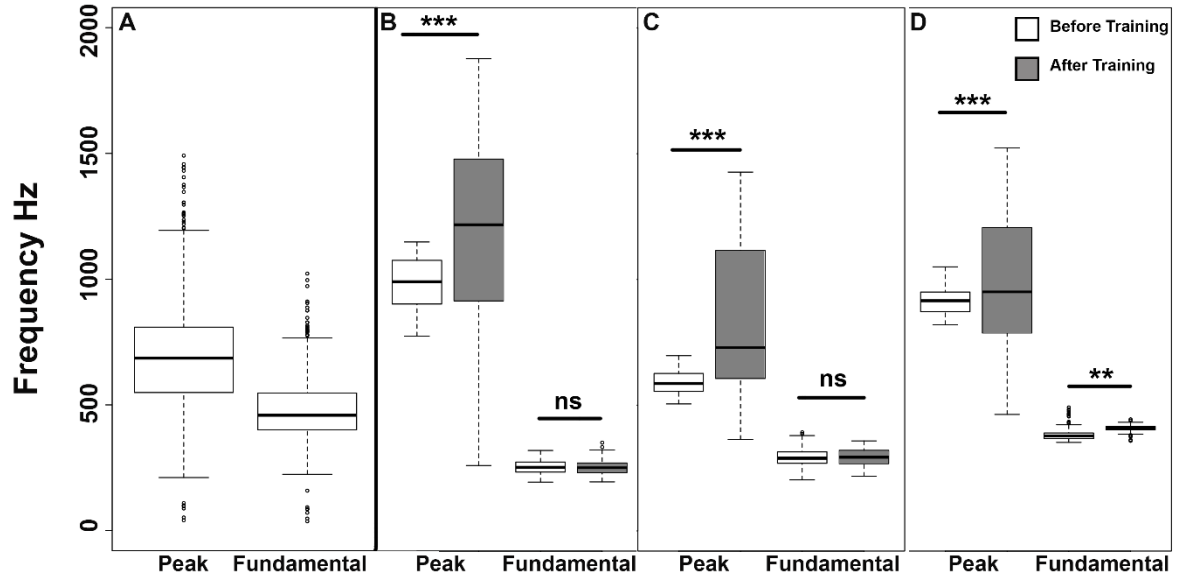

Figure 3. Tukey's boxplots of the peak and fundamental frequencies of the seal calls.

The variance in peak and fundamental frequency before training was compared to after training using Levene's test. If not significant, there was no change in variance after training. If significant, the seal's changed their calls outside of their previous repertoire variance after training.

(A) Peak and fundamental frequency of 1,859 calls recorded from 34 wild seals.

(B) Peak and fundamental frequency of Animal A, from 250 calls before and 250 calls after training.

(C) Peak and fundamental frequency of Animal B, from 250 calls before and 250 calls after training.

(D) Peak and fundamental frequency of Animal C, from 250 calls before and 250 calls after training.

Stansbury & Janik (2019) *Current Biology* 29: xxx-xxx, DOI 10.1016/j.cub.2019.05.071

Table 1. Mantel test results for all test responses and the first time novel stimuli were presented. Significant results are in bold. Animals were tested on different acoustic parameters; Animal A was tested for matching fundamental and peak frequency parameters, while Animals B and C were tested for matching formant frequencies (see also Table S1, Figure S1 and Videos S1-S4).

		Animal A		Animal B		Animal C	
	Variable	Mantel R	p-value	Mantel R	p-value	Mantel R	p-value
All Test Responses	Overall	0.292	**< 0.001**	0.247	**0.004**	0.355	**< 0.001**
	# Notes	0.928	**< 0.001**	0.946	**< 0.001**	0.929	**< 0.001**
	Fundamental Frequency	0.018	0.483	-	-	-	-
	∆ Fundamental Frequency	-0.002	0.580	-	-	-	-
	Peak Frequency	0.041	**0.011**	-	-	-	-
	∆ Peak Frequency	0.151	**< 0.001**	-	-	-	-
	Formant 1	-	-	-0.016	0.845	-0.082	1.000
	Formant 2	-	-	0.226	**< 0.001**	0.336	**< 0.001**
	Formant 3	-	-	0.057	**< 0.001**	-0.002	0.568
	Difference Formant 1 to 2	-	-	0.436	**< 0.001**	0.588	**< 0.001**
	Difference Formant 2 to 3	-	-	0.043	**< 0.001**	0.025	**0.004**
First Responses	Overall	0.417	**< 0.001**	0.399	**< 0.001**	0.358	**< 0.001**
	# Notes	0.948	**< 0.001**	0.393	**< 0.001**	0.260	**< 0.001**
	Fundamental Frequency	0.051	0.160	-	-	-	-
	∆ Fundamental Frequency	0.067	0.154	-	-	-	-
	Peak Frequency	0.115	**0.002**	-	-	-	-
	∆ Peak Frequency	0.220	**0.009**	-	-	-	-
	Formant 1	-	-	-0.034	0.936	-0.103	1.000
	Formant 2	-	-	0.276	**< 0.001**	0.273	**< 0.001**
	Formant 3	-	-	0.052	**0.006**	0.013	0.235
	Difference Formant 1 to 2	-	-	0.468	**< 0.001**	0.579	**< 0.001**
	Difference Formant 2 to 3	-	-	0.104	**< 0.001**	0.004	0.364

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further requests for resources should be directed and will be fulfilled by the Lead Contact, Vincent M. Janik

(vj@st-andrews.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Three juvenile grey seals (*Halichoerus grypus*; two females, Zola: tag numbers 73254/55, and Janice: tag numbers 73849/50, and one male, Gandalf: tag numbers 73885/86) were the subjects of this study. In the main text, Zola is referred to as 'animal A', Gandalf as 'animal B', and Janice 'animal C'. The seals were born on the Isle of May (Firth of Forth, Scotland) in November 2011 (Zola) and 2012 (Janice and Gandalf). Post weaning (approximately three weeks old) the seals were transported to the marine mammal facility at the University of St Andrews. The study animals were housed with other juvenile grey seals in three enclosures. All three enclosures were usually accessible to all animals and consisted of dry, cement haul out areas and pools of varying size (one large rectangular pool (42 x 6 x 2.5 m) and two circular pools (3 x 5 x 2 m)). The seals were fed a varied diet of several fish species. Animals were randomly allocated to experimental groups Training and testing of the seals occurred up to five days a week for 12 months. During training and testing

each seal was by itself in one of the smaller enclosures. Animals had been trained to move between enclosures so that they could be isolated for experiments. Seals were released back into the wild after 1 year of initial capture. This research was conducted under Home Office license number 60/3303. All experiments conformed to the relevant regulatory standards.

METHOD DETAILS

Training and testing summary.

To teach animals the experimental paradigm, all were initially trained to match the moan call type from their own repertoire following methods in [45]. In the second phase of training, animals heard digitally modified model calls to see how far animals could change acoustic parameters. For animal A, moan duration was kept constant while changing the pitch, allowing the relative frequency structure of the call to remain intact while linearly shifting the entire frequency spectrum of the call. We used training sequences of up to three calls, consisting of every possible combination of three different peak frequencies (880, 1046, and 1174 Hz). The seal was reinforced for matching both the number of calls and the change in frequency. During test trials we used sequences of up to 10 calls and varied the peak frequency over an octave (698-2093 Hz).

Animals B and C were presented with a new set of training sounds which varied in where energy was distributed across frequencies to correspond with formants from five of the cardinal vowels; [ɑ], [e], [i], [Ɔ] and [u]. The vowels were produced by a native North American English speaker but were digitally altered to have the same

average fundamental frequency of the seals' normal calls. Thus, these were novel calls that the seals had not produced before and that are not part of a grey seal's repertoire. To confirm this we also investigated the use of peak frequencies in 1859 calls of 34 wild naïve grey seals of similar age. Once animals B and C reached a success criterion of 80% in copying single vowel sounds, we tested them presenting the same signals in randomized combinations of up to three vowels.

Training Procedure.

All behaviours were trained using operant conditioning and positive reinforcement. A large portion of the seals' daily diet, consisting primarily of herring (*Clupea harengus*) and sprat (*Clupea sprattus*), was used as reinforcement. Correct responses were reinforced with fish, while incorrect responses resulted in a three to five second least reinforcing stimulus (LRS), a 'time-out' during which the trainer made no response, before continuing training.

All sessions took place on land and were voluntary. The seals always had access to the water, and if the animal refused to leave the water for a session, training was ended and their diet was free fed to them at the end of the day. Initially the seals were reinforced for making any sound, which progressed until they were only reinforced for vocalizing when stationed out of the water. At this point a hand cue was introduced, and the seals were reinforced for making any sound when cued in addition to staying silent when the hand cue was not present (i.e. the cue was under stimulus control).

Zola (Animal A) had then been trained to discriminate between seal call types by producing a growl when hearing a growl and producing a moan when hearing a moan. We used both her own calls and those of unknown seals as stimuli in this training. These call types have been previously well documented; growls are noisy calls with a bandwidth of up to 20 kHz, while moans are periodic calls with a harmonic structure and bandwidth rarely exceeding 5 kHz [45].

Once Zola consistently matched call type, playbacks shifted to a digitally altered stimulus. We used a moan recorded from the animal as a basis for our stimuli. The final moan stimulus always had the same duration (0.5 seconds), inter-call interval (0.1 seconds), and amplitude (70 dB re 20 µPa rms) but was digitally manipulated using Adobe Audition 2.0 to vary in number and frequency. We did not alter the natural waveform of the seal's call in any other way. The stimuli's frequency was changed using the 'pitch shifter' function, which keeps the duration of the call constant while moving the pitch of the call. This allowed the relative frequency structure of the call to remain intact while linearly shifting the frequency spectrum of the call.

Peak frequency was changed in integer steps corresponding to the musical scale nearest the seal's mean peak frequency and extending more than one standard deviation based on a sample of 100 calls (mean frequency 1015 Hz, SD ± 89.27). Thus, Zola was presented with 880, 1046, and 1175 Hz calls (corresponding with musical notes A5, C6, and D6). Shifting the peak frequency additionally changed the fundamental frequency of the signals (180, 210 and 245 Hz respectively). During each playback the seal was played one to three calls, consisting of every possible combination of these three frequencies. To enable us to reinforce performance

consistently without allowing the animal to produce the correct number of calls by stopping when reinforced, we assumed that the animal's response had finished if there was a gap of one second after her last call. This interval was chosen as it was considerably longer than the mean inter-call interval of a sample of 50 calls from her multi-call performance (mean = 0.27 seconds, SD ± 0.17). The seal was only reinforced if it matched both the number of calls and the change in frequency of calls. Rather than reinforcing an absolute match in frequency, reinforcement was based on the direction of change in peak frequency between successive stimuli such that if the change in frequency was at least 60 Hz in the correct direction (either increased or decreased as in the model, based on what was easy to judge visually on the computer screen during training and the peak frequency difference between test signals), then the seal was reinforced. If the signal consisted of only one call, then the seal was reinforced for responding with only one call regardless of frequency.

The training procedure for the seals tested in 2013, Gandalf (Animal B) and Janice (Animal C), was similar to that used in 2012, with some changes. While in 2012 the seal stationed spontaneously at a set location at the side of the pool, in 2013 both seals would position themselves near the trainer, directly next to the enclosure's fence. To keep the seals approximately one meter from the speaker, a physical station (a ball on which the seal positioned its chest) was introduced. If the seal was not at that station, its responses were not reinforced.

In initial training several different moans between 0.3 and 0.5 seconds in duration were used, with only one moan being played during each playback. Training stimuli were composed of novel sets of ten to twenty calls used per session and were

changed every two to three sessions. At this stage of training, the seals were reinforced for producing any single moan in response to the single stimulus per playback.

Once the animals were successful in this task, (in May 2013) both seals were trained with a single moan that always had the same duration (1.0 second) and amplitude (70 dB re 20 µPa rms) but varied in frequency and number of repetitions. Peak frequency was changed in integer steps corresponding to the musical scale nearest the seal's mean peak frequency and extending more than one standard deviation in both directions (for Janice 915 Hz, SD ± 54.08 and Gandalf 577 Hz, SD ± 31.32). Thus, for Janice signals of 783, 987 and 1175 Hz (corresponding with musical notes G5, B5, and D6) were used while for Gandalf stimuli had frequencies of 493, 587, and 698 Hz (B4, D5 and F5). During each playback the seal was played between one and three moans per playback with any combination of the three frequencies. Just as in 2012, the seals were only rewarded when producing the correct number of calls and changed the frequency of their calls in the correct direction as in the sample.

Once the seals both had five consecutive sessions with at least 80% correct responses, they were then presented with a new set of training stimuli. These calls had constant duration (0.6 seconds), amplitude (70 dB re 20 µPa rms) and fundamental frequency. The fundamental frequency was chosen based on each seal's average for a sample of 100 calls; for Janice 380 Hz (mean 378, SD ± 33.79) and Gandalf 190 Hz (mean 192, SD ± 39.02). The calls varied by sound spectrum (i.e. where energy was distributed across frequencies), to correspond with formants from five of the cardinal vowels; [ɑ], [e], [i], [Ɔ] and [u]. These specific vowels were

chosen as they are produced using variable mouth, lip and tongue positions and they were easily identifiable by human listeners. The vowels were produced by a native North American English speaker (recorded with sampling rate 96 kHz, 24-bit) and then digitally altered in Adobe Audition to have a mean fundamental frequency similar to that of the seals. The seals were presented with one sound per playback and reinforced for responding with the correct number (i.e. one call) and formant frequencies. Once the seals had five consecutive sessions with 80% correct trials, they were then presented with multiple calls. At this time the seals were presented with up to five vowels (inter-call interval 0.05 seconds) per trial, always in the same order (i.e. vowels were always played in the following sequence: [ɑ], [e], [i], [Ɔ] and [u]) in addition to being played individually. Thus, nine training signals were used at this point.

Testing Procedure.

Each trial was initiated by the seal moving into position, stationed at the side of the pool with head facing the speaker. A sound was played, and the seal was judged to have finished calling when no additional sounds were produced for more than one second. The seal's response was evaluated using a real time spectrographic display in Audacity. If correct, the seal was reinforced with pieces of fish. If incorrect, a LRS of three to five seconds occurred before beginning the next trial.

Test stimuli were created using one of the seals previously recorded vocalizations digitally altered to vary in number and frequency, for Zola, or number and formant frequencies, for Gandalf and Janice. While during training Zola was only exposed to up to three frequencies (880, 1046, and 1174 Hz), test stimuli extended over an

octave (ranging from 698 to 2093 Hz, corresponding to musical notes F5 to C7). Throughout testing 15 sequences consisting of different combinations of up to ten of these notes were used. Each sequence was presented with every possible number of notes (i.e. for a three note melody, the first note was presented alone, the first and second note alone, and the full three note melody). Thus, in total there were 92 stimuli used throughout testing.

For Zola, stimulus order was arbitrarily randomized by the researcher with one exception. In order to prevent frustration, if the seal responded incorrectly in a trial, the next stimulus presented was one the researcher subjectively thought the seal would be more likely to succeed at. This would either be a shorter sequence of the same melody or an alternative melody the seal had higher average accuracy with. This procedure resulted in stimuli being played an unequal number of times. Every stimulus was played at least once (mean 2.924, SD ± 2.765).

For Janice and Gandalf, test stimuli consisted of the same five cardinal vowels used in training ([ɑ], [e], [i], [Ɔ] and [u]). While in training these vowels were only presented in one order, during testing they were presented in randomized combinations of up to three vowels. Every possible combination was tested, resulting in 155 different stimuli, each played three times throughout testing. Thus, each seal was tested with 465 trials.

Acoustic recordings.

Recordings were collected using a Sennheiser MKH 416 P48 directional microphone (frequency response 40-20,000 Hz, sensitivity at 1 kHz 25 mV/Pa +/-1 dB) and Edirol FA-66 external sound card (sampling rate 96 kHz, 24-bit) onto a laptop computer.

Sounds were played with the same system through an external Skytec active speaker (frequency response 32-22,000 Hz). Sounds were simultaneously played, recorded and spectrographically monitored in real time using the program Audacity 1.3 (sampling rate 96 kHz, 24-bit; Audacity Team, 2012). Weather permitting, sessions were concurrently video recorded using a Sony HDR CX250E video camera.

Additional recordings of a wild population of greys seals on the Isle of May, Scotland, were taken in November and December 2011 using a Sennheiser MKH 416 P48 directional microphone (frequency response 40 Hz to 20 kHz, sensitivity at 1 kHz 25 mV/Pa +/-1 dB) on a Marantz Pro Solid-state recorder PMD671 (sampling rate 96 kHz, 24 bit, high pass filter @ 100 Hz). We recorded a total of 1,859 moans from 34 seal pups during 175 hours (aged from birth through weaning) of observation. These served as a baseline to assess the natural repertoire of juvenile grey seal moans.

QUANTIFICATION AND STATISTICAL ANALYSIS

The number of sounds, fundamental, and peak frequency were measured using Avisoft-Saslab Pro 5.02.04 sonogram software. Formant frequencies were measured using Praat version 5.3.51. See Table S1 for definitions of the test parameters. For Praat analysis, the settings were kept the same for each individual animal, and only varied slightly to account for the higher frequency sounds produced by animal C. The following settings were used and proofed useful for analysing seal sounds: number of formants: 6, window length: 0.01, Dynamic range: 40 dB, Max formant: 6,000 Hz

(animal B) 6,500 (animal C). Formant frequencies were measured every 25 ms, resulting in 25 measured points per call.

In total, n=3,725 sounds were used for analysis (n=1,425 from Animal A, n=1,125 from Animal B, and n=1,175 from Animal C). The similarity between the signals played and the seals' response was evaluated using distance matrices and the Mantel statistic. As each seal was presented with different signals, statistics were run for each animal individually. "Gower" distance was measured using the pairwise dissimilarities between sounds calculated by the daisy function in the cluster package for R 1.15.2. The daisy function was chosen as it allows for distances to be calculated for multivariate mixed data. Thus, we were able to obtain a measure of overall similarity using all measured parameters in addition to testing each parameter individually. This was done for all responses combined. In addition, a subset of the data was also analyzed. The seals response to the first presentation of each stimuli was analyzed separately to measure how accurate the seal's performance was for novel calls.

Two separate matrices were calculated; one for the signals played and one for the seals' response. Vectors between matrices were aligned such that each point matched the signal played with the seals corresponding response. The Mantel test was then used to measure the association between the signal and response matrices with Pearson's product-moment correlation coefficient [57] using the mantel function in the vegan package for R 2.0-10. Matrix correlations were compared to chance by random reallocation of the matrix elements using 1,000 permutations, which results in a 95% confidence level [58]. To avoid autocorrelation, elements were allocated such that call sequences were kept intact [58]. This test produces two

statistics, the Mantel R and p-value. The Mantel R is a value between -1 and 1, where 0 shows no correlation, -1 shows a negative correlation, and 1 a positive correlation between matrices. The p-value statistic compares the correlation to random chance, based on the randomized permutations.

Throughout training, the seals were reinforced for producing sounds with increasingly variable frequencies. Thus, it would be anticipated that at the start of training the seals produced more stereotyped calls and by the end of training produced more variable calls. To check if this occurred, the peak frequency of the first 100 calls made at the start of training was compared to the last 100 calls recorded from each seal. The variance between groups was compared with Levene's test using the package lawstat for R 2.4.1. Levene's test was chosen because the peak frequencies were not normally distributed.

LEGENDS FOR SUPPLEMENTARY VIDEOS

Video S1. Zola Scale. Related to Figure 1. The grey seal Zola (animal A) listens to a scale and then copies it.

Video S2. Zola tune. Related to Table 1. The grey seal Zola (animal A) listens to a tune and then copies it.

Video S3. Janice vowels e-a-a. Related to Figure 1. The grey seal Janice (animal C) listens to the vowel sequence e-a-a and then copies it. Janice had a distinct voice compared to the other seals, using higher frequencies and sounding noisier than others. This was the case for her usual seal calls as well.

Video S4. Gandalf vowels a-u-u. Related to Figure 1. The grey seal Gandalf (animal B) listens to the vowel sequence a-u-u and then copies it.

REFERENCES

1. Janik, V.M., and Slater, P.J.B. (2000). The different roles of social learning in vocal communication. Anim. Behav. *60*, 1-11.
2. Petkov, C.I., and Jarvis, E.D. (2012). Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. Front. Evol. Neurosci. *4*, 12.
3. Catchpole, C.K., and Slater, P.J.B. (2008). Bird song: biological themes and variations, 2nd edition (Cambridge: Cambridge University Press).
4. Janik, V.M., and Slater, P.J.B. (1997). Vocal learning in mammals. Adv. Study Behav. *26*, 59-99.
5. Fitch, W.T. (2010). The evolution of language (Cambridge: Cambridge University Press).
6. Russell, J.L., McIntyre, J.M., Hopkins, W.D., and Taglialatela, J.P. (2013). Vocal learning of a communicative signal in captive chimpanzees, *Pan troglodytes*. Brain Lang. *127*, 520-525.
7. Lameira, A.R., Hardus, M.E., Kowalsky, B., de Vries, H., Spruijt, B.M., Sterck, E.H.M., Shumaker, R.W., and Wich, S.A. (2013). Orangutan (Pongo spp.) whistling and implications for the emergence of an open-ended call repertoire: A replication and extension. J. Acoust. Soc. Am. *134*, 2326-2335.
8. Takahashi, D.Y., Fenley, A.R., Teramoto, Y., Narayanan, D.Z., Borjon, J.I., Holmes, P., and Ghazanfar, A.A. (2015). The developmental dynamics of marmoset monkey vocal production. Science *349*, 734-738.
9. Watson, S.K., Townsend, S.W., Schel, A.M., Wilke, C., Wallace, E.K., Cheng, L., West, V., and Slocombe, K.E. (2015). Vocal learning in the functionally referential food grunts of chimpanzees. Curr. Biol. *25*, 495-499.

10. Takahashi, D.Y., Liao, D.A., and Ghazanfar, A.A. (2017). Vocal learning via social reinforcement by infant marmoset monkeys. Curr. Biol. *27*, 1844-1852.

11. Fischer, J., Wheeler, B.C., and Higham, J.P. (2015). Is there any evidence for vocal learning in chimpanzee food calls? Curr. Biol. *25*, R1028-R1029.

12. Balsby, T.J.S., Momberg, J.V., and Dabelsteen, T. (2012). Vocal imitation in parrots allows addressing of specific individuals in a dynamic communication network. Plos One *7*, e49747.

13. Wanker, R., Sugama, Y., and Prinage, S. (2005). Vocal labelling of family members in spectacled parrotlets, *Forpus conspicillatus*. Anim. Behav. *70*, 111-118.

14. Knörnschild, M., Nagy, M., Metz, M., Mayer, F., and von Helversen, O. (2010). Complex vocal imitation during ontogeny in a bat. Biol. Lett. *6*, 156-159.

15. Boughman, J.W. (1998). Vocal learning by greater spear-nosed bats. Proc. R. Soc. Lond. B *265*, 227-233.

16. Janik, V.M. (2013). Cognitive skills in bottlenose dolphin communication. Trends Cogn. Sci. *17*, 157-159.

17. Noad, M.J., Cato, D.H., Bryden, M.M., Jenner, M.N., and Jenner, K.C.S. (2001). Cultural revolution in whale song. Nature *408*, 537.

18. Richards, D.G., Wolz, J.P., and Herman, L.M. (1984). Vocal mimicry of computer-generated sounds and vocal labeling of objects by a bottlenosed dolphin, *Tursiops truncatus*. J. Comp. Psychol. *98*, 10-28.

19. Poole, J.H., Tyack, P.L., Stoeger-Horwath, A.S., and Watwood, S. (2005). Elephants prove capable of vocal learning. Nature *434*, 455-456.

20. Flower, T. (2011). Fork-tailed drongos use deceptive mimicked alarm calls to steal food. Proc. R. Soc. Lond. B *278*, 1548–1555.

21. Todt, D. (1975). Social learning of vocal patterns and modes of their application in grey parrots (*Psittacus erithacus*). Z Tierpsychol *39*, 178-188.

22. Klatt, D.H., and Stefanski, R.A. (1974). How does a mynah bird imitate human speech? J. Acoust. Soc. Am. *55*, 822-832.

23. West, M.J., Stroud, N., and King, A.P. (1983). Mimicry of the human voice by European starlings: the role of social interaction. Wilson Bull. *95*, 635-640.

24. Stoeger, A.S., Mietchen, D., Oh, S., de Silva, S., Herbst, C.T., Kwon, S., and Fitch, W.T. (2012). An Asian elephant imitates human speech. Curr. Biol. *22*, 2144-2148.

25. Ralls, K., Fiorelli, P., and Gish, S. (1985). Vocalizations and vocal mimicry in captive harbor seals, *Phoca vitulina*. Can. J. Zool. *63*, 1050-1056.

26. Fitch, W.T., and Hauser, M.D. (2010). Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. In Acoustic communication, A.M. Simmons, A.N. Popper and R.R. Fay, eds. (New York: Springer), pp. 65-137.

27. Reber, S.A., Nishimura, T., Janisch, J., Robertson, M., and Fitch, W.T. (2015). A Chinese alligator in heliox: formant frequencies in a crocodilian. J. Exp. Biol. *218*, 2442-2447.

28. Taylor, A.M., Charlton, B.D., and Reby, D. (2018). Vocal production by terrestrial mammals: source, filter, and function. In Vertebrate sound production and acoustic communication, R.A. Suthers, W.T. Fitch, R.R. Fay and A.N. Popper, eds. (Cham: Springer), pp. 229-259.

29. Riede, T., and Zuberbühler, K. (2003). The relationship between acoustic structure and semantic information in Diana monkey alarm vocalization. J. Acoust. Soc. Am. *114*, 1132-1142.

30. Townsend, S.W., Charlton, B.D., and Manser, M.B. (2014). Acoustic cues to identity and predator context in meerkat barks. Anim. Behav. *94*, 143-149.

31. Farago, T., Pongracz, P., Range, F., Viranyi, Z., and Miklosi, A. (2010). 'The bone is mine': affective and referential aspects of dog growls. Anim. Behav. *79*, 917-925.

32. Fitch, W.T., and Fritz, J.B. (2006). Rhesus macaques spontaneously perceive formants in conspecific vocalizations. J. Acoust. Soc. Am. *120*, 2132-2141.

33. Ghazanfar, A.A., Turesson, H.K., Maier, J.X., van Dinther, R., Patterson, R.D., and Logothetis, N.K. (2007). Vocal-tract resonances as indexical cues in rhesus monkeys. Curr. Biol. *17*, 425-430.

34. Ravignani, A., Fitch, W.T., Hanke, F.D., Heinrich, T., Hurgitsch, B., Kotz, S.A., Scharff, C., Stoeger, A.S., and de Boer, B. (2016). What pinnipeds have to say about human speech, music and the evolution of rhythm. Front. Neurosci. *10*, 274.

35. McCulloch, S., and Boness, D.J. (2000). Mother-pup vocal recognition in the grey seal (*Halichoerus grypus*) of Sable Island, Nova Scotia, Canada. J. Zool. *251*, 449-455.

36. Fagot, J., and De Lillo, C. (2011). A comparative study of working memory: immediate serial spatial recall in baboons (*Papio papio*) and humans. Neuropsychologia *49*, 3870-3880.

37. Pembrook, R.G. (1986). Interference of the Transcription Process and Other Selected Variables on Perception and Memory during Melodic Dictation. J. Res. Music Educ. *34*, 238-261.

38. Moore, B.R. (1996). The evolution of imitative learning. In Social learning in animals: the roots of culture, C.M. Heyes and B.G. Galef, eds. (San Diego, CA: Academic Press), pp. 245-265.

39. Abramson, J.Z., Hernandez-Lloreda, V., Garcia, L., Colmenares, F., Aboitiz, F., and Call, J. (2018). Imitation of novel conspecific and human speech sounds in the killer whale (*Orcinus orca*). Proc. R. Soc. Lond. B *285*, 20172171.

40. Lilly, J.C. (1965). Vocal mimicry in *Tursiops*: ability to match numbers and durations of human vocal bursts. Science *147*, 300-301.

41. Murayama, T., Iijima, S., Katsumata, H., and Arai, K. (2014). Vocal imitation of human speech, synthetic sounds and beluga sounds, by a beluga (*Delphinapterus leucas*). Int. J. Comp. Psychol. *27*, 369-384.

42. Ridgway, S., Carder, D., Jeffries, M., and Todd, M. (2012). Spontaneous human speech mimicry by a cetacean. Curr. Biol. *22*, R860-R861.

43. Janik, V.M. (2014). Cetacean vocal learning and communication. Curr. Opin. Neurobiol. *28*, 60-65.

44. Reidenberg, J.S., and Laitman, J.T. (2010). Generation of sound in marine mammals. In Handbook of mammalian vocalization: an integrative neuroscience approach, S.M. Brudzynski, ed. (London: Academic Press), pp. 451-465.

45. Shapiro, A.D., Slater, P.J.B., and Janik, V.M. (2004). Call usage learning in gray seals (*Halichoerus grypus*). J. Comp. Psychol. *118*, 447-454.

46. Stansbury, A.L., de Freitas, M., Wu, G.-M., and Janik, V.M. (2015). Can a gray seal (*Halichoerus grypus*) generalize call classes? . J. Comp. Psychol. *129*, 412-420.

47. Byrne, R.W. (2002). Imitation of novel complex actions: What does the evidence from animals mean? Adv. Study Behav. *31*, 77-105.

48. Carouso-Peck, S., and Goldstein, M.H. (2019). Female social feedback reveals non-imitative mechanisms of vocal learning in zebra finches. Curr. Biol. *29*, 631-636.

49. Insley, S.J., Philips, A.V., and Charrier, I. (2003). A review of social recognition in pinnipeds. Aquat. Mamm. *29*, 181-201.

50. Insley, S.J. (2000). Long-term vocal recognition in the northern fur seal. Nature *406*, 404-405.

51. Bjørgesæter, A., Ugland, K.I., and Bjørge, A. (2004). Geographic variation and acoustic structure of the underwater vocalization of harbor seal (*Phoca vitulina*) in Norway,Sweden and Scotland. J. Acoust. Soc. Am. *116*, 2459-2468.

52. Terhune, J.M., and Dell'Apa, A. (2006). Stereotyped Calling Patterns of a Male Weddell Seal (*Leptonychotes weddellii*). Aquat. Mamm. *32*, 175-181.

53. Cleator, H.J., Stirling, I., and Smith, T.G. (1989). Underwater vocalizations of the bearded seal (*Erignathus barbatus*). Can. J. Zool. *67*, 1900-1910.

54. Sjare, B., Stirling, I., and Spencer, C. (2003). Structural variation in the songs of Atlantic walruses breeding in the Canadian High Arctic. Aquat. Mamm. *29*, 297-318.

55. Reichmuth, C., and Casey, C. (2014). Vocal learning in seals, sea lions, and walruses. Curr. Opin. Neurobiol. *28*, 66-71.

56. Sanvito, S., Galimberti, F., and Miller, E.H. (2007). Observational evidence of vocal learning in southern elephant seals: a longitudinal study. Ethology *113*, 137-146.

57. Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Res. *27*, 209-220.

58. Manly, B.F.J. (1997). Randomization, bootstrap and Monte Carlo methods in biology (London: Chapman and Hall).

Stansbury & Janik (2019) *Current Biology* 29: xxx-xxx, DOI 10.1016/j.cub.2019.05.071

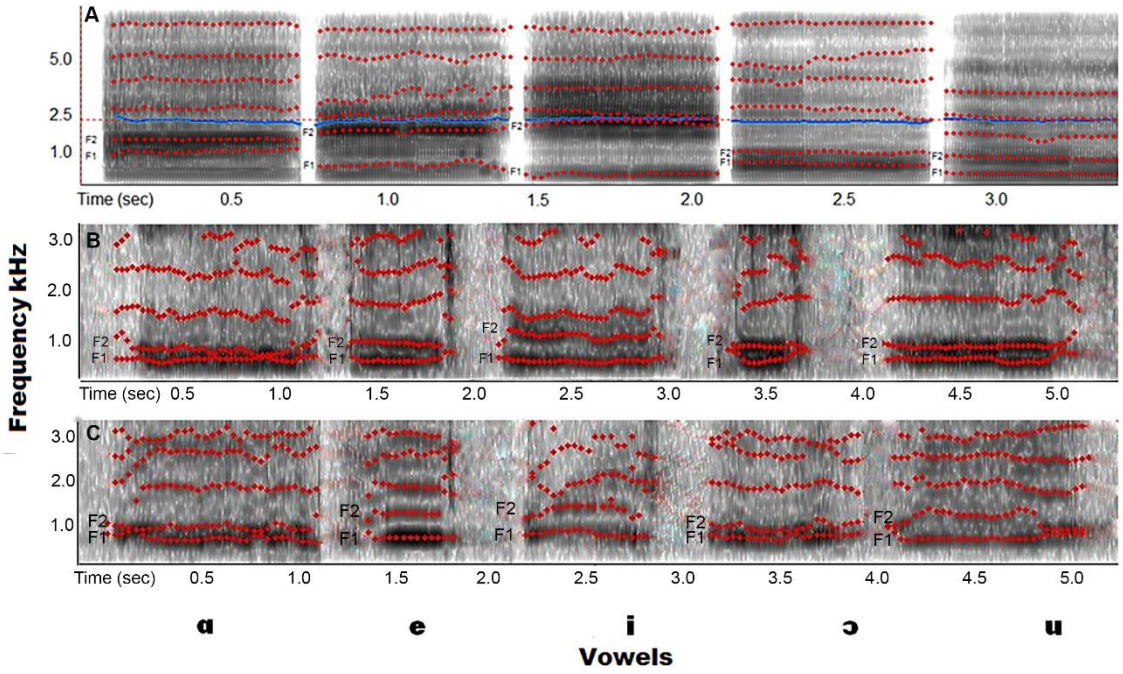

Figure S1. (A) Test stimuli for animals B and C. (B) Example responses by animal B. (C) Example responses by animal C. Related to Figure 1 and Table 1. Spectrogram displays produced in Praat. Individual vowels are listed along the x-axis, with the red dots corresponding to formant frequencies. The first (F1) and second (F2) formant frequency is shown next to each vowel. The blue line in (A) corresponds to the sounds overall pitch, as calculated using Boersma formula (approximately 2.5 kHz).

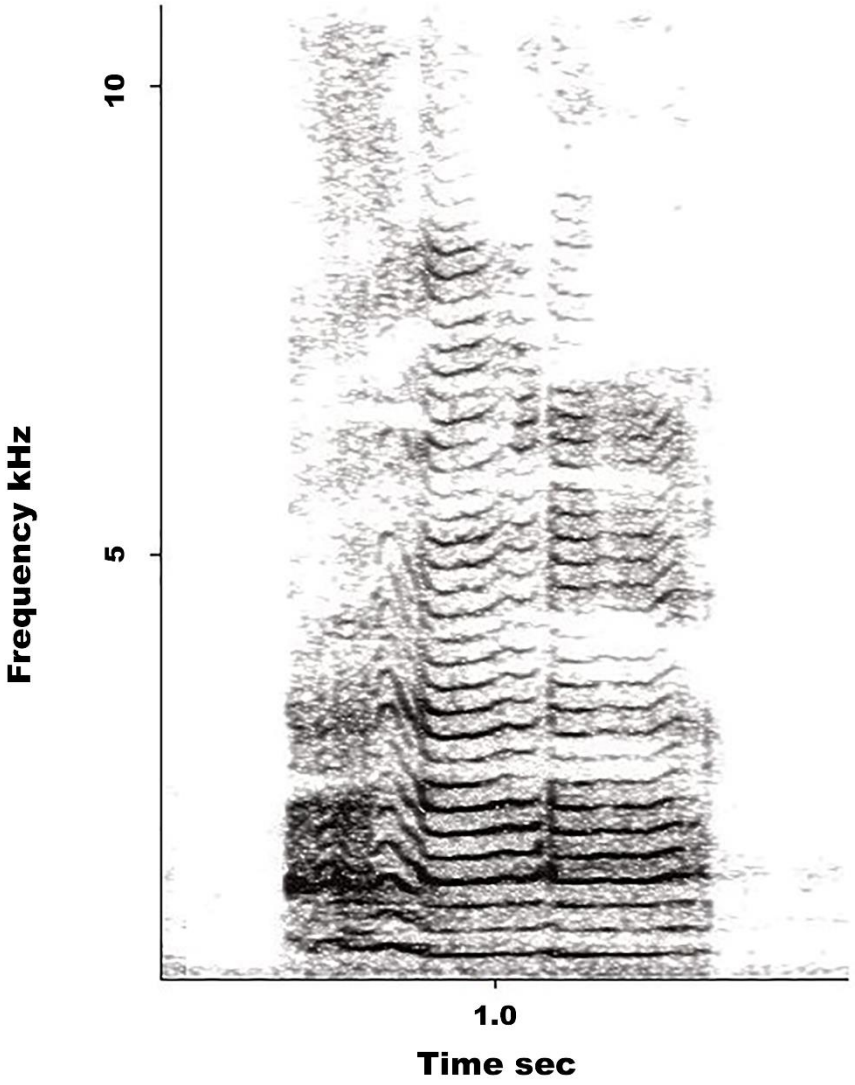

Figure S2. Example of grey seal moan. Related to Figure 1. Grey seal moans are tonal calls with harmonic structure. Example taken from animal A prior to imitation training. Spectrogram created in Avisoft-SASlab Pro (FFT size 2048, frequency resolution 47 Hz, time resolution, 10.7 ms, weighting function: hamming window, window width 100%).

Parameter	Definition
# of sounds	Number of individual sounds without a break in frequency of more than 5 ms within -35 dB of the maximum spectrum peak, ending when 1 second passed without any additional calls being made.
Fundamental Frequency	Frequency in Hz of the lowest integer multiple of amplitude peaks in a harmonic call. Measured every 5 ms and averaged across the call.
∆ Fundamental Frequency	The difference in fundamental frequency (Hz) between consecutive calls. Only measured in multiple call responses, with no measure taken for the first response in a sequence.
Peak Frequency	The frequency with the highest amplitude measured every 5 ms and averaged across the call.
∆ Peak Frequency	The difference in peak frequency (Hz) between consecutive calls. Only measured in multiple call responses, with no measure taken for the first response in a sequence.
Formant 1	Automatically measured in Praat as the first peak above the fundamental in the calls spectrum. Measured every 5 ms and averaged across the call.
Formant 2	Automatically measured in Praat as the second peak in frequency in the calls spectrum. Measured every 5 ms and averaged across the call.
Formant 3	Automatically measured in Praat as the third peak in frequency in the calls spectrum. Measured every 5 ms and averaged across the call.
Difference Formant 1-2	The difference in frequency (Hz) between the first and second formant.
Difference Formant 2-3	The difference in frequency (Hz) between the second and third formant.

Table S1. Acoustic parameters used during analysis and their definitions. Related to Figure 2 and Table 1.